



LEARNING AS PARAMETER OPTIMIZATION / 2

The ERM optimization problem is:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \mathcal{R}_{\text{emp}}(\theta).$$

For a **(global) minimum** $\hat{\theta}$ it obviously holds that

$$\forall \theta \in \Theta : \quad \mathcal{R}_{\text{emp}}(\hat{\theta}) \leq \mathcal{R}_{\text{emp}}(\theta).$$

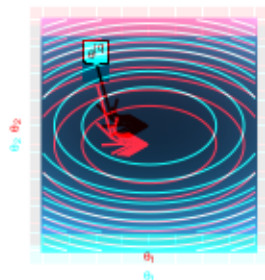
This does not imply that $\hat{\theta}$ is unique.

Which kind of numerical technique is reasonable for this problem strongly depends on model and parameter structure (continuous params? uni-modal $\mathcal{R}_{\text{emp}}(\theta)$?). Here, we will only discuss very simple scenarios.

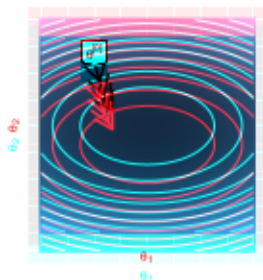


GRADIENT DESCENT - LEARNING RATE

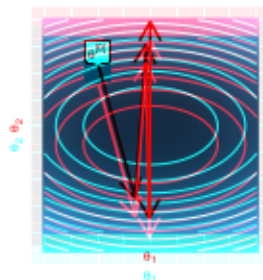
- The negative gradient is a direction that looks locally promising to reduce \mathcal{R}_{emp} .
- Hence it weights components higher in which \mathcal{R}_{emp} decreases more.
- However, the length of $-\frac{\partial}{\partial \theta} \mathcal{R}_{\text{emp}}$ measures only the local decrease rate, i.e., there are no guarantees that we will not go "too far".
- We use a learning rate α to scale the step length in each iteration. Too much can lead to overstepping and no converge, too low leads to slow convergence.
- Usually, a simple constant rate or rate-decrease mechanisms to enforce local convergence are used



good convergence for α_1
good convergence for α_2



poor convergence for $\alpha_2 (< \alpha_1)$
poor convergence for $\alpha_3 (< \alpha_1)$



no convergence for $\alpha_3(> \alpha_1)$
no convergence for $\alpha_3(> \alpha_1)$

